Pre-training and Evaluating Transformer-based Language Models for Icelandic

Jón Friðrik Daðason, Hrafn Loftsson
Department of Computer Science, Reykjavik University, Iceland
{jond19, hrafn}@ru.is

Introduction

- We present four new Transformer-based language models for Icelandic
- The models are evaluated on a selection of four tasks
- We compare our results against two previously published language models as well as previous state-of-the-art models

Pre-training

We pre-train small and medium-sized **ELECTRA** and **ConvBERT** models for 1M steps on the **Icelandic Gigaword Corpus** (IGC), which contains 1.69B tokens from various domains. Each model uses a byte-pair encoding tokenizer with a vocabulary size of 32k

Model	Params	Time (TPU v3-8)
ELECTRA-Small	14M	8 hrs.
ConvBERT-Small	13M	15 hrs.
ELECTRA-Base	110M	10 days
ConvBERT-Base	107M	13 days

Fine-tuning

- We evaluate the models on four tasks: **part-of-speech tagging**, **named entity recognition**, **dependency parsing** and **extractive text summarization**
- We compare our results against two other previously published language models
  - **IceBERT-IGC**, a RoBERTa-Base model (124M param.) pre-trained on the IGC
  - **mBERT**, a multilingual Bert-Base model (179M param.) pre-trained on Wikipedia
- We also include results obtained by previous state-of-the-art models

❑ Part-of-Speech Tagging

For part-of-speech tagging, we evaluate the models on the **MIM-GOLD** corpus, which consists of 1M tokens annotated with POS tags. We compare our results against **ABLTagger**, a BiLSTM model which previously obtained state-of-the-art results on MIM-GOLD. We report tagging accuracy using 10-fold cross validation. We find that ConvBERT-Base obtains the best results, reducing the error rate of ABLTagger by 54%.

Model	Accuracy
ABLTagger	95.15%
mBERT	96.38%
ELECTRA-Small	96.84%
ConvBERT-Small	96.88%
IceBERT-IGC	97.37%
ELECTRA-Base	97.72%
ConvBERT-Base	97.75%

❑ Named Entity Recognition

For NER, we evaluate the models on **MIM-GOLD-NER**, a version of MIM-GOLD which has been annotated with named entities. We also evaluate a BiLSTM model which previously obtained state-of-the-art results on the corpus. We report entity-level F-scores using 10-fold cross validation. ConvBERT-Base obtains the best results, reducing the error rate of the BiLSTM model by 55% and of ConvBERT-Small by 26%.

Model	F-score
BiLSTM	87.07%
ELECTRA-Small	91.23%
mBERT	91.31%
ConvBERT-Small	92.03%
IceBERT-IGC	93.04%
ELECTRA-Base	93.75%
ConvBERT-Base	94.14%

❑ Dependency Parsing

For dependency parsing, we evaluate the models on the **Icelandic Parsed Historical Corpus** (IcePaHC), which contains 1M tokens from documents dating from the 11th to the 21st century, that were manually annotated with constituents. We use **IcePaHC-UD**, a version of the corpus that has been converted to the Universal Dependencies format. We report labeled attachment scores (LAS), using 70% of the corpus of training, 15% for validation and 15% for testing. We find that ConvBERT-Base again obtains the best results.

Model	LAS
mBERT	82.94%
ConvBERT-Small	84.75%
ELECTRA-Small	84.90%
IceBERT-IGC	85.04%
ELECTRA-Base	86.20%
ConvBERT-Base	86.50%

❑ Extractive Text Summarization

For text summarization, we evaluated the models on **IceSum**, a corpus containing 1,000 Icelandic news articles that have been manually annotated with extractive summaries. We also report results for the **Lead** baseline, which generates a summary from the first few sentences of a document, which has historically been difficult to outperform when summarizing news articles. We also report results for a BiLSTM model with a sequence-to-sequence based sentence extractor, which previously obtained state-of-the-art performance on IceSum. We report ROUGE-2 recall scores, averaged over 5 runs, using 70% of the corpus for training, 15% for validation and 15% for testing. We find that three models outperform the Lead baseline by a statistically significant margin: Seq2Seq, ELECTRA-Base and ConvBERT-Base

Model	ROUGE-2
mBERT	69.09
Lede-100	69.14
IceBERT-IGC	69.14
ELECTRA-Small	69.29
ConvBERT-Small	69.36
Seq2Seq	70.42
ELECTRA-Base	71.04
ConvBERT-Base	71.09
Oracle	89.48

Conclusions

- We publish four monolingual language models for Icelandic, which have been made available on the HuggingFace model repository
  - https://huggingface.co/jonfd
- The larger models obtain significantly better results than both their smaller versions and previous state-of-the-art models
  - The smaller models obtain good performance overall, but fail to outperform the baseline for extractive text summarization
- The multilingual mBERT model obtains similar performance to the smaller monolingual models, despite being 13 times larger

Acknowledgements

- This project was funded by the Language Technology Programme for Icelandic 2019-2023. The programme, which is managed and coordinated by Almannarómur, is funded by the Icelandic Ministry of Education, Science and Culture.
- This research was supported with Cloud TPUs from Google's TPU Research Cloud (TRC).